Improving English Pronunciation: An Automated Instructional Approach

Sugata Mitra,[a] James Tooley,[b]
Parimala Inamdar,[c] and Pauline Dixon[d]

a.
sugatam@niit.com
Centre for Research in
Cognitive Systems
NIIT Limited
Synergy Building
IIT Campus
New Delhi, India

b.
j.n.tooley@ncl.ac.uk
University of Newcastle
Newcastle Upon Tyne, England

c.
parimalai@niit.com
Centre for Research in
Cognitive Systems
NIIT Limited
Synergy Building
IIT Campus
New Delhi, India

d.
pauline.dixon@ncl.ac.uk
University of Newcastle
Newcastle Upon Tyne, England

Abstract

This paper describes an experiment in which groups of children attempted to improve their English pronunciation using an English-language learning software, some English films, and a speech-to-text software engine. The experiment was designed to examine two hypotheses. The first is that speech-to-text software, trained in an appropriate voice, can act as an evaluator of accent and clarity of speech as well as help learners acquire a standard way of speaking. The second is that groups of children can operate a computer and improve their pronunciation and clarity of speech, on their own, with no intervention from teachers. The results of the experiment are positive and point to a possible new pedagogy.

Introduction

It has been assumed that speech-to-text (STT) programs cannot distinguish good pronunciation and have not been used for this purpose (Goodwin-Jones 2000). However, we propose that STT programs can be used instead of human listeners to evaluate the quality of pronunciation by comparing against a standard accent and pronunciation. The proposition is based on a recent patent application (Mitra and Ghatak 2002).

To investigate this hypothesis, we need to discover whether an STT program can judge pronunciation as well as a human can. If an STT program were demonstrated to be as good as a human judge of pronunciation, it could be suggested that human speakers, particularly children, may be able to improve their pronunciation on their own, using such a program rather than requiring a human teacher. This hypothesis is inspired by, and uses findings of, previous research that shows that groups of children can learn to use computers to self-instruct without teachers (Mitra 1988, 2003; Mitra and Rana 2001). Popularly referred to as the "hole-in-the-wall," a series of experiments showed that groups of children (8–13 years old) were able to self-instruct in using computers for playing games, surfing the Internet, painting, and other activities, irrespective of their social, linguistic, or ethnic backgrounds. Working in groups and unsupervised, free access was important to the self-instructional process.

This paper is made up of four parts. First, the reasons there is desirability for better pronunciation are examined. Second, the research method and results are described, setting out how, where, and when the data were collected. Third, the findings of the research are discussed, and finally the paper closes with a conclusion and a vision for the way forward.

The Environment and the Problem

Hyderabad is a large city in southern India and, like many such cities in India, contains sprawling slum areas. These areas in Hyderabad contain a large number of small private schools, sometimes as many as 10 in a square kilometre. These schools are filled to capacity with children whose parents pay substantial amounts for education (in comparison to their incomes) in spite of the fact that there are several free schools in the vicinity operated by the government. The single most important reason the slum parents send their children to these private schools is the English language.

There are 17 languages recognized by the United Nations, and more than 700 dialects related to these languages, spoken in India. Hindi is the national language, and English is a common "bridging" language that is used everywhere.

As a consequence of its colonial past, people who speak English are generally considered more suitable for most jobs than people who do not. Although this may not be a happy situation, it is one of the main reasons India scores over other Southeast Asian nations in the software industry. India is the second largest exporter of software in the world after the United States, and its success is often attributed to the ability of its industry and its people to deal with the English language.

The ability to speak in English can determine the living standards and occupations of most Indians. It is for this reason that the private schools in Hyderabad prosper.

Although these schools teach English with a reasonable effectiveness, they suffer from a severe mother-tongue influence (MTI). The products of such schools, and, indeed, any school in India, can be fluent in English but would often speak the language with an accent that is incomprehensible anywhere in the world. The reason for this is because teachers in such schools have a strong MTI themselves, which their students copy, and the problem perpetuates itself. It is in this context that the experiment in this study was conducted.

Why Is Better Pronunciation Desirable?

People from all countries are now working and living in a globalized environment where communication from and to almost anywhere in the world may occur practically instantaneously. Labor mobility and the existence of international employment opportunities have heightened the need to communicate and to be understood. The most recent of such services are in the information technology (IT) enabled services (ITES) area. These services use communication technology to leverage the (often cheaper) work force of one country to service the requirements of another. For example, the cost of taking orders on the telephone is cheaper in India than in the United States because of lower salaries for telephone operators. As a result, many U.S. companies have set up call centers in India, where telephone operators take orders for American goods from American customers, who are unaware that the conversation they are having is with an Indian located in India. English is generally regarded as the language that can provide this communication universality, so much so that parents know how important it is for their children to master the English language for them to succeed. Moreover, in developing countries it has been found that not only do the elite and middle classes aspire for their children to acquire and master English language skills but also parents from poor slum areas now place great importance on their children attaining the ability to read, write, and speak in English. In the private schools in the low-income areas of Hyderabad, India, for instance, it is shown that the single most important element for parents choosing a private school is that it teaches the entire curriculum in the English vernacular (Tooley and Dixon Forthcoming). However, such ambitions are thwarted because good-quality teachers (and, in particular, native English speakers) are not available to provide, inter alia, good English pronunciation language skills to their students. As a result, even the best students from the system have a strong MTI on their English pronunciation and their speech is only barely understandable. This, in turn, affects their employability and social status severely, leading to underemployment and, sometimes, crime, poverty, and associated urban social problems.

The need to be understood for cultural as well as ITES purposes is the reason behind the undertaking of the experiment. We decided to conduct the experiment in an area of great educational need, where untrained teachers were catering to children whose parents were illiterate and where English was not widely spoken although greatly desired. The experiment sets out to explore two hypotheses:

- Hypothesis 1: The objective measurement of words correctly recognized by an STT program closely parallels subjective human judgments of pronunciation.
- Hypothesis 2: Given suitable software and hardware, children, and in particular disadvantaged children, can instruct themselves, without requiring a teacher, in the improvement of their English pronunciation.

Method

Location and Sample Size

The experiment took place in Hyderabad, India, between September 2002 and January 2003. A computer was placed in a private, unaided school (Peace High School, Edi Bazaar), which caters to very low-income families, situated in a slum area of the city. The school fees are between 75 rupees to 150 rupees per month (about US$1.50 to US$3 per month). The school serves children of auto-rickshaw drivers, market traders, and service workers, with approximate monthly income ranging from 1,000 rupees to 3,000 rupees per month. Most families have more than one child; hence, the school fees can exceed 25% of the monthly incomes. A sample of 16 children from different classes in the school, between 12 and 16 years old, were chosen at random to participate in the study. The school is an English medium school; the entire curriculum is taught in English. The mother tongue of the children is either Urdu or Telegu. The children could already read and speak English at variable standards of competency.

Equipment and Apparatus: Hardware and Software

The school was provided with a computer. The multi-media computer was a Pentium P4, 1.8 GHz, 256 Mb RAM, with microphone and speakers. The specialized software used was "Ellis Kids" and "Dragon Naturally Speaking," using the Windows XP operating system. The system was installed in a quiet part of the school. This ensured that the sounds received by the microphone were only those of the learners and not external noise. STT programs are often confused in noisy environments.

Ellis

Ellis (www.planetellis.com) is one of the leading English language learning software programs in the market, released in 1992. Ellis Kids is its specialized children's package. It teaches vocabulary, listening, grammar, and communication skills through multimedia inputs such as video, audio, and text. Learning is tested through multiple-choice questions; the students are also able to test their pronunciation by recording a few words of their speech and comparing it with a standard English spoken voice. Ellis is not equipped with any text-to-speech capability and is not capable of providing the user with any feedback on the quality of the learner's pronunciation. Indeed, there is no software at this time that can do so. However, the focus of Ellis is the whole of language learning, of which pronunciation is a small part.

Dragon

Dragon Naturally Speaking (www.dragonsystems.com) is an STT engine. Dragon can be trained to recognize, interpret, and convert to text a particular user's speech. This is done by comparing the audio input with a stored profile. To create a profile, a user has to read several passages provided by the software into the computer's microphone. These readings are then converted into a profile for the user, about 3 Mb in size, which needs to be selected before speech can be entered into the word processing program. Naturally, Dragon will best recognize those users' speech in whose voices it has been trained. This experiment used the Dragon STT program uniquely as a learning and testing tool for English pronunciation. Four profiles of speech were created in Dragon: Pauline, English female, James, English male, Sugata, Indian male, Parimala, Indian female. These were used as pronunciation standards against which students practiced and were measured. Students read passages into Dragon using any one of these profiles. For the final measurements, all male voices were measured against the profile Sugata, and all female voices were measured against the profile Parimala.

Films

Four films were also installed in the computer as additional environmental inputs for spoken English. These were *The Sound of Music, My Fair Lady, Guns of Navarone,* and *The King and I.* There are two reasons for introducing these films into the experiments. First, the students have no reference to how they should correctly pronounce an English word, although Dragon may tell them that their pronunciation is incorrect. By watching 4 hours of films, we expected them to have heard most of the common

English words as spoken by native English speakers. The second reason for selecting these films was to break the monotony of practicing a few passages from a school reader everyday. The films were chosen because all, except *Guns of Navarone*, had something to do with education and children or young people. *Guns of Navarone* was chosen because it contains the kind of action that young people seem to enjoy.

Procedure

Students, consisting of eight girls and eight boys, were randomly organized into four groups of four and each group was instructed to spend a total of 3 hours per week with the system (Figure 1). A timetable was made out for this purpose and given to the students.

The students were then given a demonstration of Ellis and Dragon software and were told the names and locations of the movies stored in the computer. They were given no instructions other than that they should try to improve their English pronunciation by reading passages from their school English textbooks into Dragon and to try to make Dragon recognize as much of their readings as possible. They were told to use Ellis if they wished, or to watch any film of their choice on the computer, if bored. It was suggested to each group that they should help each other and collectively organize what they would do with the computer when it was their turn to use it. There was to be no organized adult intervention after this point, except in case of any equipment failure and related maintenance activity. We relied on the earlier results (Mitra 1988, 2003; Mitra and Rana 2001) that groups of children can instruct themselves without teachers if given adequate computing resources. In effect, the children were left with a clear objective (to alter their speech until understood by the Dragon STT program) and a demonstration of the resources available (i.e., computer, Dragon, Ellis, and movies). They were asked to organize themselves to meet this objective and they

Figure 1. Children at Peace High School, Edi Bazaar, Hyderabad, India, working on their computer.

readily agreed to try. Indeed, they agreed with great enthusiasm.

Measurements

The experiment was started in September 2003. Each month, a researcher visited the school and asked all 16 children to read passages from a standard English textbook into the Dragon STT program. From October, they took measurements of the students' reading, comparing the resultant text produced by Dragon with the correct text, and calculated the percentage of correctly identified words. Each passage was of approximately 100 words. The text produced by Dragon was done using girls' readings compared against the Parimala profile, whereas the boys' readings were compared against the Sugata profile. Dragon performs poorly when a female profile is used to judge a male voice and vice versa. This is because of obvious frequency differences in male and female voices. It is interesting to note that, for very young children of either gender, a female profile should be used for the same reason.

Each month, starting in October, a new passage was added to the one in the previous month; that is, the children read one passage in September and October, two in November, and so on until January 2003, when they read four passages. This was done so that their pronunciation in the first passage, which they read four times in the 4 months since October, could be compared with their pronunciation in the fourth passage, which they read for the first time in January.

It should be mentioned that there were interruptions in the school calendar because of examinations and holidays during the experiment, which reduced the learners' time spent at the computer. We estimate that each child spent between 10 and 15 hours on the computer during this period. Measurements were taken at monthly intervals from September 2002 to January 2003; that is, five measurements in total. In the first month, the students used Ellis only, and they began using Dragon in October. This was unavoidable because the Dragon program was unavailable before October.

A video clip was made of each child reading each passage in each of these monthly sessions. The video clips were recorded so that human judges would be able to watch and rank the pronunciations of the children reading the passages. At this point it is important to explain why we chose to record the readings using video, and not just audio, as the medium.

The first hypothesis, as described earlier, is to determine whether an STT program, such as Dragon, would score and rank the pronunciations of human speakers such that these scores and ranks are similar to those by human judges. That is, the judgment of the STT program and of human judges should show concordance. Because Dragon is "judging" pronunciation only by "listening" to the speakers, it would seem natural that the judging process for humans should also use only the audio from each speaker. After all, it is possible that a human judge would score differently if he or she were to be looking at a speaker or only listening to a recording of a speaker's voice during the judging process.

Some reflection shows that this reasoning is incorrect in the current context. First, we can find no evidence that the human judgment of pronunciation is affected by whether the judge is only listening to a voice or looking at the speaker. Second, an STT program such as Dragon can neither "listen" nor "judge" in any manner that can be remotely comparable to the process followed by humans. Humans process speech and vision in a way that is, as far as current understanding goes, different from the way computers do (e.g., see Bhatnagar et al. 2002). Dragon does not listen; it receives a stream of bits (zeros and ones) and processes them mathematically against another stored array of bits—the "profile." It is therefore futile to try to give the humans and an STT program the "same" input, in this case, audio. What is more relevant is that the STT program should be shown to behave as an instrument that can measure pronunciation in a way such that its measurements reflect the same values as those produced by human, subjective judgment. A measuring instrument does not necessarily have to measure the same parameter as a human being to come to similar conclusions. For example, an EEG machine can detect an epileptic attack by measuring electrical signals from the patient's brain, whereas a human being can detect the same attack by other, subjective audiovisual means.

In the present context, what we wish to study in the children is their clarity and pronunciation in spoken English as judged by other humans. In society, such judgments are made more often in face-to-face situations than otherwise. Indeed, in the present

context of the children of Hyderabad, India, the judgment of the effectiveness of their spoken English will be almost entirely done by others who are in their presence. Although it is conceivable that such a judgment may be made over telephone or some other media where the speaker is not present visually, it is not considered an important or relevant factor for the present experiment.

We have therefore presented the human judges with video clippings of the children reading, to be closest to the social context in which their English will be judged in real life. Dragon, on the other hand, has been presented with audio bit streams of these readings (in the 44 KHz, mono *.wav* format), which is the only input it can accept.

Sometimes, some of the children were absent for a session and their data could not be collected. Although this is undesirable, it did not affect more than 10% of the planned readings. The data at the end of the experiment consisted of 16 children's reading of passages 1, 2, 3, and 4 over 5 months, for a total of 160 readings, 160 resultant text files from Dragon, and 160 video clips.

Analysis of data

Analysis of data was done at the end of the experiment. At this time, the children were ranked for each of the readings in each of the months in order of the percentage of words correctly identified by Dragon, for each passage.

The video clips were labeled and randomized, and viewed by four human judges, who are the authors of this paper. Each judge, while viewing a video clip, was unaware of when the recording had been made. Because the process of scoring was done only once, at the end of the experiment, it was important that the judges be unaware of the date of each recording. The judges then scored each video clip in terms of their *subjective* assessment, on a score of 1 to 10 (10 high), of how well each child performed in English pronunciation, including clarity of speech and accent. Eventually, the scores obtained were used to rank the children according to each judge.

It is important to establish first that the human judges are, themselves, in agreement with each other before proceeding to match their rankings with that of Dragon. Kendall's coefficient of concordance (Siegel 1956) was calculated for the four judges for each of the 5 months for this purpose. Kendall's coefficient, *W,* is considered more effective for smaller samples (less than 30) than Cohen's kappa for measuring concordance.

The same coefficient (*W*) was also calculated with Dragon acting as a "fifth judge"; that is, the ranking obtained from the process using Dragon was added to those obtained from the four human judges.

Control

A control group is needed to verify the second hypothesis. However, it was difficult to obtain a control group at the beginning of the experiment. Any child tested for pronunciation would want to be a part of the experimental "course," as they perceived it. Because the equipment would be left in charge of the children and no adult control would be imposed on them, it would be impossible to monitor if any child from the control group was also joining the experimental group. Based on advice from the school principal, we decided to not measure any child other than those chosen for the experimental group. These children were instructed not to let other children use the computer, an instruction they followed meticulously. As a result, no control group was tested at the beginning of the experiment, and this creates limitations to the generalizability of the results. The following method was adopted to rectify this situation.

At the end of the experimental period (January 2003), 16 children were selected at random from the same sections of the same school as those from which the experimental groups had been selected. These children were aware of the fact that some of their peers were working on a computer, but they had no other information about what they were doing. Neither did this new group of children have any exposure to computers or to any other teaching-learning method other than that used traditionally by the school. Therefore, the only difference between the new (control) group and our experimental group was that the latter had worked on the computer as described earlier for 5 months. The control group was asked to read out three passages from a standard English textbook, and a video clip was produced for each reading session.

It was now necessary to establish whether the scores of the children using Dragon and Ellis for 5 months were significantly different from those of the children who had not. To do this, we decided to

use four new judges to remove any bias, because the previous judges were already familiar with the experimental group, having visited the school many times. In this context, a judge has to be someone whose English pronunciation is of an acceptable standard. All judges chosen were such that their own recognition rates in Dragon were 60% or more. Incidentally, the best STT programs, including Dragon, generally show a recognition rate of 85% or less, even when used by the same person who has trained it.

Four new judges were selected and presented with a collection of 48 video clips: 16 from the experimental group's readings of September 2002, 16 from the experimental groups readings from January 2003, and 16 from the control groups readings of January 2003. The clips were presented to the judges in random order to ensure that they had no way of determining either the nature of the group (experimental or control) or the time of the video recording. Each reading was given a score from 1 to 10 (10 high) for the same *subjective* parameters as used earlier by each judge. The average scores for each group of 16 readings were calculated, the readings were converted to ranks, and the Kendall's coefficient of concordance was calculated for the four new judges.

Results

Table 1 shows the results for the Kendall's coefficient of concordance. This table reveals two interesting factors. First, the human judges, although judging subjectively, were in very strong concordance, with *W* ranging from 0.79 in October to 0.94 in January, with $p < 0.001$ in every case. Second, the addition of the fifth judge, Dragon, did drop the coefficient of concordance, but it remained at a significant degree of concordance, ranging from 0.69 in November to 0.81 in January, with $p < 0.001$ again.

Figure 2 then shows the average scores of the four human judges for the entire group of 16 children, compared with the average Dragon score for the entire group, over the five-month period (Dragon has no score for the first month). Two interesting factors are revealed again. First, the self-instructional process using the software dramatically improved the results for reading: a 117% increase for the human judges in terms of their subjective judgments of pronunciation, and a 79% increase for the fifth judge, Dragon, in terms of percentage of words recognized. Second, that the curves are closely correlated suggests again that the method of objectively measuring words recognized by Dragon parallels the human method of subjectively judging pronunciation clarity and accent. In Figure 2, it appears as though the human judge scores are consistently higher than those of Dragon. This is not important, as the two scores can be normalized to the same scale. We have not done so because it would then be visually difficult to see the two sets of scores, human and Dragon, separately.

In the control part of the experiment as described earlier, the Kendall coefficient for the four new judges was found to be 0.71, again showing good agreement among them. The average scores for the experimental group's readings in September 2002 and the control group's readings in January 2003, shown in Table 2, were found to be exactly equal at 30% each. The average score for the experimental group's readings in January 2003 were found to be

*Table 1. Kendall's Coefficient of Concordance (*W*)*

Data for 16 Students Reading 3 Passages	***W* for 4 Human Judges**				***W* for 4 Human Judges and 1 STT Judge**			
	W	χ^2	df	*p*	*W*	χ^2	df	*p*
September	0.83	49.5	15	0.001				
October	0.79	31.7	10	0.001	0.70	34.8	10	0.001
November	0.80	44.7	14	0.001	0.69	48.2	14	0.001
December	0.84	43.8	13	0.001	0.73	47.2	13	0.001
January	0.94	52.5	14	0.001	0.81	56.8	14	0.001

Note: χ^2 = chi-square; df = degrees of freedom, p *= the probability of* W *being nonsignificant; STT = speech-to-text.*

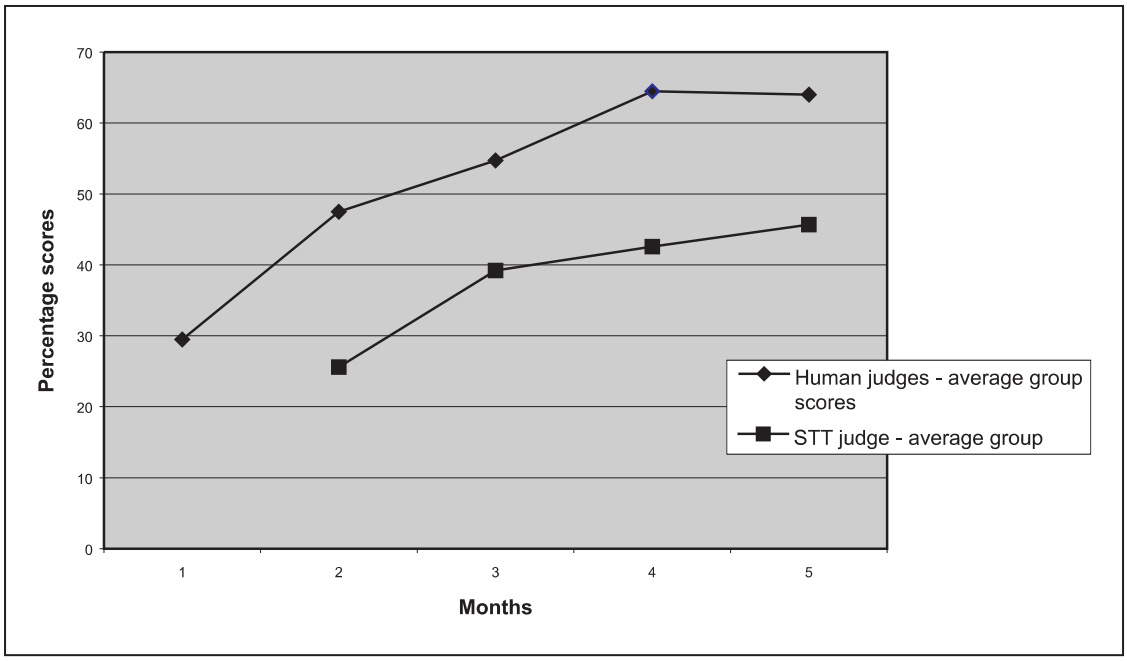

Figure 2. Improvements in pronunciation as measured by human judges and by a speech-to-text program.

Table 2. Independent Comparison of Experimental and Control Group Scores

Description of Video Clips	Average Percent Scores from 4 Judges
Experimental Group, September 2002	30
Experimental Group, January 2003	72
Control Group, January 2003	30

significantly higher at 72%. This is shown in Figure 3.

It may be noticed that STT (Dragon) scores were not used in this part of the measurement. This is because the purpose of these measurements was to compare the results of the experimental and control groups alone. Therefore, correlating these with STT results was not required in this part of the measurement.

Discussion and conclusions

The results are interesting for both of the hypotheses. Regarding the first, the high concordance values shown in Table 1 seem to indicate unequivocally that the method of using the objective measure of percentage of words correctly recognized by Dragon closely mimics the human process of subjectively judging children's pronunciation. This seems to be an important result, with possibly significant implications for the automation of the judgment of pronunciation. Another intriguing possibility is that of automatically detecting speakers with a certain accent.

The procedure followed with the control group and four independent judges shows clearly that the control group's scores at the end of the experiment in January match the experimental group's scores at the beginning of the experiment in September. In other words, the control group children, with no exposure to the experimental apparatus and procedure, continued to have pronunciations similar to those of the experimental group children before they began their interaction with the computer and software. The experimental group, on the other hand, showed significant improvement after the experimental period.

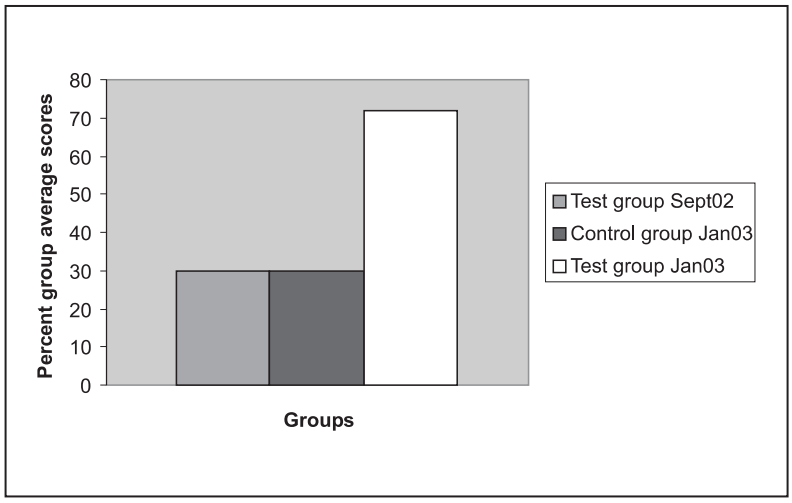

Figure 3. Scores of experimental and control groups in different months as marked by four independent judges.

Ideally, we should have scored the control group at the beginning of the experiment as well. However, because this could not be done and because it is very unlikely that the control group's scores at beginning of the experimental period could have been *higher* than at the end of the period, we propose on the basis of Figure 3 that the improvements in pronunciation observed in the experimental group of children were caused by the experimental procedure. The fact that this is supported by an independent, randomized judging process further supports this contention.

It seems highly unlikely that such improvements in pronunciation could have been effected by other things in the school or slum environment. The school principal corroborated that he had noticed significant improvements in the language ability of these children but not others. Certainly, we can say that when groups of children are given the appropriate resources, they *can* improve their pronunciation with minimal intervention from adults. Further work is required to see to what extent their improvement was based on which parts of the software packages used (Ellis, Dragon, or, indeed, *My Fair Lady!*) and the degree to which these improvements exceed those that might be expected over time without such interventions.

Yet another possibility emerges from these results. It appears that, using an STT system, children can acquire any kind of accent, because the accent acquired is similar to the one in which the system is trained. ■

Acknowledgments

We would like to thank Mr. Wajid, principal of Peace High School, Edi Bazaar, Hyderabad, India, for his support for this experiment. Detailed comments from one of the referees were invaluable and resulted in a dramatic improvement in the paper.

References

Bhatnagar, G., S. Mehta, and S. Mitra, eds. 2002. *Introduction to Multimedia Systems.* San Diego, Calif.: Academic Press.

Eastment, D. 1998. *ELT and the New Technology: The Next Five Years.* Retrieved from www.eastment.com.

Ehsani, F. and E. Knodt. 1998. "Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm." *Language Learning & Technology* 2:45–60.

Goodwin-Jones, B. 2000. "Emerging Technologies: Speech Technologies for Language Learning." *Language Learning & Technology* 3:6–9.

Mitra, S. 1988. "A Computer-Assisted Learning Strategy for Computer Literacy Programmes." Paper presented at the annual convention of the

All-India Association for Educational Technology, Goa, India.

Mitra, S. 2003. "Minimally Invasive Education: A Progress Report on the 'Hole-in-the-wall' Experiments." *British Journal of Educational Technology* 34:367–371.

Mitra, S. and R. Ghatak. 2002. "An apparatus for measuring clarity of spoken English," Indian patent application number 1159/DEL/2002, November 18.

Mitra, S. and V. Rana. 2001. "Children and the Internet: Experiments with Minimally Invasive Education in India." *The British Journal of Educational Technology* 32:221–232.

Siegel, S. 1956. *Nonparametric Statistics for the Behavioural Sciences.* New York: McGraw-Hill.

Stark, D. G. 1997. *Hal's Legacy: 2001's Computer as Dream and Reality.* Cambridge, MA: MIT Press.

Tooley, J. and P. Dixon. Forthcoming. "Providing Education for the World's Poor: A Case Study of the Private Sector in India," in B. Davies and J. West-Burnham, eds., *Handbook of Educational Leadership and Management.* Victoria, Australia: Pearson Longman.